DOI: 10.11992/tis.201703047
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20170702.1548.038.html

多层递阶融合模糊特征映射的模糊 C 均值聚类算法

鲍国强[1,2], 应文豪[3], 蒋亦樟[1,2], 张英[1,2], 王骏[1,2], 王士同[1,2]
（1. 江南大学 数字媒体学院, 江苏 无锡 214122; 2. 江苏省媒体设计与软件技术重点实验室, 江苏 无锡 214122; 3. 常熟理工学院 计算机科学与工程学院, 江苏 常熟 215500）

摘　要: 针对复杂非线性数据的无监督学习问题, 提出一种新型的映射方式来有效提高算法对复杂非线性数据的学习能力。以 TSK 模糊系统的规则前件学习为基础, 提出一种新型的模糊特征映射新方法。接着, 针对映射之后的数据维度过大问题, 引入多层递阶融合的概念, 进一步提出基于多层递阶融合的模糊特征映射新方法, 从而有效避免了因单层模糊特征映射之后特征维数过高而导致的数据混乱和冗余的问题。最后与模糊 C 均值算法相结合, 提出基于多层递阶融合模糊特征映射的模糊 C 均值聚类算法。实验研究表明, 文中算法相比于经典模糊聚类方法, 有着更加优越、稳定的性能。

关键词: Takagi-Sugeno-Kang (TSK) 模糊系统; 主成分分析 (PCA); 无监督学习; 模糊 C 均值聚类
中图分类号: TP181　　**文献标志码:** A　　**文章编号:** 1673−4785(2018)04−0594−08

中文引用格式: 鲍国强, 应文豪, 蒋亦樟, 等. 多层递阶融合模糊特征映射的模糊 C 均值聚类算法[J]. 智能系统学报, 2018, 13(4): 594–601.
英文引用格式: BAO Guoqiang, YING Wenhao, JIANG Yizhang, et al. Fuzzy C-means clustering algorithm for multilayered hierarchical fusion fuzzy feature mapping[J]. CAAI transactions on intelligent systems, 2018, 13(4): 594–601.

Fuzzy C-means clustering algorithm for multilayered hierarchical fusion fuzzy feature mapping

BAO Guoqiang[1,2], YING Wenhao[3], JIANG Yizhang[1,2], ZHANG Ying[1,2], WANG Jun[1,2], WANG Shitong[1,2]
(1. School of Digital Media, Jiangnan University, Wuxi 214122, China; 2. Jiangsu Key Laboratory of Media Design and Software Technology, Wuxi 214122, China; 3. School of Computer Science and Engineering, Changshu Institute of Technology, Changshu 215500, China)

Abstract: In this paper, we propose a novel feature mapping technique called multilayer hierarchical fusion fuzzy feature mapping for the unsupervised learning of complex nonlinear data and combine it with the classical fuzzy C-means clustering. Based on the regular antecedent learning of the Takagi-Sugeno-Kang (TSK) fuzzy system, we first propose a novel fuzzy feature mapping method. Then, to address big data dimensions by fuzzy feature mapping, we propose a fuzzy feature mapping mechanism based on multilayer hierarchical fusion. This mechanism combines fuzzy feature mapping with principal component analysis (PCA), thereby avoiding the data confusion and redundancy caused by the high dimensionality of single-layer fuzzy feature mapping. Finally, we develop a novel FCM clustering algorithm based on multilayered hierarchical fusion feature mapping. The experimental results show that, in comparison with classical fuzzy clustering methods, the performance of the proposed algorithm is superior and more stable.

Keywords: Takagi-Sugeno-Kang (TSK) fuzzy system; principal component analysis (PCA); unsupervised learning; fuzzy C-means clustering

收稿日期: 2017−03−30.　**网络出版日期:** 2017−07−02.
基金项目: 国家自然科学基金项目 (61300151); 江苏省自然科学基金项目 (BK20160187, BK20161268, BK20151299); 江苏省产学研前瞻联合研究计划项目 (BY2015043-03).
通信作者: 王骏. E-mail: wangjun_sytu@hotmail.com.

近年来, 面向复杂非线性数据的模糊聚类问题得到了研究人员的广泛关注[1-6]。在无监督学习环境中为了提高复杂非线性数据的可分性, 一个

重要的研究思路是使用非线性映射将数据映射到高维空间中。在众多非线性映射方法中，核方法作为经典的隐性映射方法得到了广泛的应用[5-13]。研究表明，核方法通过使用核函数代替内积运算，将待分类数据隐性地映射到高维空间，从而有助于复杂非线性数据的学习。但是，核方法还存在着诸多局限性，尤其是如何针对不同的问题选择合适的核函数和相关参数，这都会影响算法的聚类效果。

模糊系统因其强大的不确定性系统建模能力、优良的可解释性和出色的泛化能力，近年来在复杂非线性数据学习问题中得到了大量的研究。在已有的经典模糊系统中，Takagi-Sugeno-Kang(TSK)[14-17]模糊系统由于其良好的解释性和简洁性得到了广泛应用。在 TSK 模糊系统中，其规则前件部分通过显性映射方式 (本文称之为模糊特征映射)，将输入数据映射到高维空间中去。从本质上讲，模糊特征映射可以视为一种特殊的非线性映射方式。基于此，本文将输入数据进行相应的非线性映射。在具体实现过程中我们发现，经模糊特征映射后的特征维数过高，这会增加计算量，同时也导致了数据的冗余。为此，本文通过引入多层递阶融合机制和主成分分析，提出新型的基于多层递阶融合的模糊特征映射新方法。并将之与经典模糊聚类技术相结合，进一步提出基于多层递阶融合模糊特征映射的模糊 C 均值聚类新方法。经实验验证，本文算法在处理复杂非线性数据时能够取得比传统模糊聚类算法更有效的聚类效果。

1　Takagi-Sugeno-Kang 模糊系统及模糊特征映射

Takagi-Sugeno-Kang 模糊系统模型[18-23]是最重要的用于建模与智能控制的模糊模型之一。对于经典的 TSK 模糊模型，最常用的模糊推理规则的定义如下：

第 k 条模糊规则：

IF

$$\boldsymbol{x}_1 \text{ is } \boldsymbol{A}_1^k \wedge \boldsymbol{x}_2 \text{ is } \boldsymbol{A}_2^k \wedge \cdots \boldsymbol{x}_d \wedge \text{ is } \boldsymbol{A}_d^k$$

THEN

$$f^k(\boldsymbol{x}) = \boldsymbol{p}_0^k + \boldsymbol{p}_1^k \boldsymbol{x}_1 + \cdots + \boldsymbol{p}_d^k \boldsymbol{x}_d, k = 1,2,\cdots,K \tag{1}$$

式中：$\boldsymbol{A}_i^k$表示输入向量 $\boldsymbol{X}$ 第 i 维特征所对应的第 k 条模糊规则的模糊子集；K 表示模糊规则数；$\wedge$为模糊合取操作。每条规则都对应输入向量 $\boldsymbol{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_d]^{\mathrm{T}}$，并且把输入空间的模糊子集 $\boldsymbol{A}^k \subset \boldsymbol{R}^d$映射到输出空间的模糊集 $f^k(\boldsymbol{X})$，其中乘算子、加算子分别作为合取和析取操作算子，加法算子作为组合算子时，TSK 模糊模型的输出可以表示为

$$y^0 = \sum_{k=1}^{K} \frac{\mu^k(\boldsymbol{X})}{\sum_{k'=1}^{K} \mu^{k'}(\boldsymbol{X})} f^k(\boldsymbol{X}) = \sum_{k=1}^{K} \tilde{\mu}^k(\boldsymbol{X}) f^k(\boldsymbol{X}) \tag{2}$$

式中：$\mu^k(\boldsymbol{X})$和$\tilde{\mu}^k(\boldsymbol{X})$分别表示为模糊集 $\boldsymbol{A}^k$相关的模糊隶属函数和归一化模糊隶属函数。这两个函数的计算公式分别为

$$\mu^k(\boldsymbol{X}) = \prod_{i=1}^{d} \mu_{A_i^k}(\boldsymbol{x}_i) \tag{3}$$

和

$$\tilde{\mu}^k(\boldsymbol{X}) = \mu^k(\boldsymbol{X}) / \sum_{k'=1}^{K} \mu^{k'}(\boldsymbol{X}) \tag{4}$$

通常采用高斯函数作为模糊隶属函数，其计算公式为

$$\mu_{A_i^k}(\boldsymbol{x}_i) = \exp\left(\frac{-(\boldsymbol{x}_i - \boldsymbol{c}_i^k)^2}{2\boldsymbol{\delta}_i^k} \right) \tag{5}$$

式中：参数$\boldsymbol{c}_i^k$和$\boldsymbol{\delta}_i^k$可以通过聚类技术或其他划分方法计算得出。通常使用模糊 C 均值 (FCM) 聚类算法进行数据集的初始划分，进而计算 $\boldsymbol{c}_i^k$和$\boldsymbol{\delta}_i^k$公式为

$$\boldsymbol{c}_i^k = \sum_{j=1}^{N} \boldsymbol{u}_{jk} \boldsymbol{x}_{ji} / \sum_{j=1}^{N} \boldsymbol{u}_{jk} \tag{6}$$

$$\boldsymbol{\delta}_i^k = h \cdot \sum_{j=1}^{N} \boldsymbol{u}_{jk} \left(\boldsymbol{x}_{ji} - \boldsymbol{c}_i^k \right)^2 / \sum_{j=1}^{N} \boldsymbol{u}_{jk} \tag{7}$$

式中：$\boldsymbol{u}_{jk}$表示通过 FCM 聚类计算出属于第 k 类的第 j 个输入数据 $\boldsymbol{X}_j = [\boldsymbol{x}_{j1}\ \boldsymbol{x}_{j2}\ \cdots\ \boldsymbol{x}_{jd}]^{\mathrm{T}}$的模糊隶属度。在这里，$h$ 是高斯函数的核带宽参数。令：

$$\boldsymbol{X}_e = \left[1\ \boldsymbol{X}^{\mathrm{T}} \right]^{\mathrm{T}} \tag{8}$$

$$\tilde{\boldsymbol{X}}^k = \tilde{\mu}^k(\boldsymbol{X}) \boldsymbol{X}_e \tag{9}$$

$$\boldsymbol{X}_g = \left[\left(\tilde{\boldsymbol{X}}^1 \right)^{\mathrm{T}} \left(\tilde{\boldsymbol{X}}^2 \right)^{\mathrm{T}} \cdots \left(\tilde{\boldsymbol{X}}^K \right)^{\mathrm{T}} \right]^{\mathrm{T}} \tag{10}$$

$$\boldsymbol{P}^k = \left[\boldsymbol{p}_0^k\ \boldsymbol{p}_1^k\ \cdots\ \boldsymbol{p}_d^k \right]^{\mathrm{T}} \tag{11}$$

$$\boldsymbol{P}_g = \left[\left(\boldsymbol{P}^1 \right)^{\mathrm{T}} \left(\boldsymbol{P}^2 \right)^{\mathrm{T}} \cdots \left(\boldsymbol{P}^K \right)^{\mathrm{T}} \right]^{\mathrm{T}} \tag{12}$$

TSK 模糊模型的训练问题转化为式 (13) 线性回归模型的参数学习问题[24]：

$$y^0 = \boldsymbol{P}_g^{\mathrm{T}} \boldsymbol{X}_g \tag{13}$$

从式 (13) 中可以观察到，输入向量经式 (8) ~ (10) 计算，可以变换为一个 $(d+1)*K$ 维的高维向量，本文中我们将这一转换过程称为模糊特征映射。与已有核方法中的隐性映射相比，模糊特征映射具有以下特点：1) 它是一种显性映射方式，

用户可以在高维特征空间中得到数据的显式表示方法；2) 模糊特征映射基于模糊规则进行构建，而模糊规则本身具有较强的可解释性；3) 输入向量经模糊特征映射后得到的高维特征向量的维数可以由模糊规则数确定，这有利于用户控制高维空间中数据的复杂程度。

2 基于多层递阶融合模糊特征映射的模糊 C 均值聚类算法

2.1 基于多层递阶融合的模糊特征映射新方法

原数据通过模糊特征映射，得到其在高维空间中的新表示。但是作为单层映射结构，会因映射后的特征维数过高使得数据变得混乱和冗余，继而影响算法后续的聚类效果。研究表明[25-26]，将单层映射结构改造为多层映射结构，可以有效地提高算法对复杂非线性数据的学习能力。为此，本文引入多层递阶融合的概念来构造新型的映射，提出基于多层递阶融合的模糊特征映射新方法 (MLHFFFM)。通过对每层模糊特征映射之后的高维特征表示进行 PCA 降维，再进行相应的信息补充，形成新的融合层，依次进入下一层的压缩融合过程，其结构如图 1 所示。

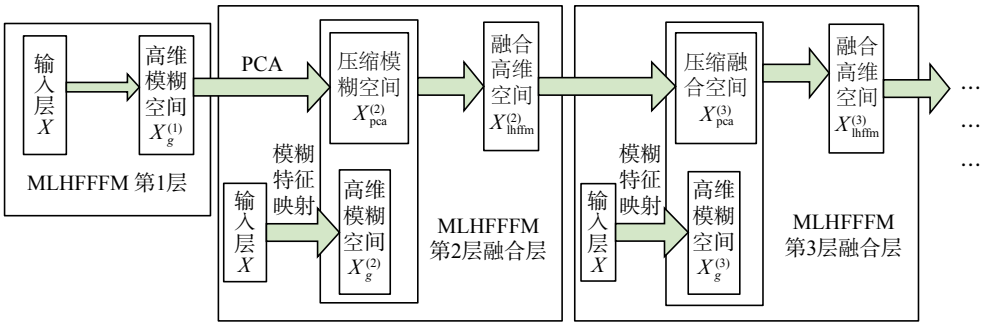

图 1　MLHFFFM 算法结构图

Fig. 1　Structure of MLHFFFM algorithm

基于多层递阶融合的模糊特征映射新方法 MLHFFFM 算法描述如下：

输入　给定一个数据集 $\boldsymbol{D}=\{\boldsymbol{X}, \boldsymbol{Y}\}$，设置初始模糊规则数 K，分层融合层数 S。

输出　经多层递阶融合后的数据矩阵 $\boldsymbol{X}_{pca}^{(S)}$。

1) 对原数据进行第一层的模糊特征映射 (初始层)

① 通过 FCM 算法计算出隶属度矩阵 $\boldsymbol{u}_{jk}$；

② 经式 (6) 和式 (7) 分别计算出对应的 $\boldsymbol{c}_i^k$ 和 $\boldsymbol{\delta}_i^k$ (i=1, 2, …, d, j=1, 2, …, n, k=1, 2, …, K)；

③ 通过高斯隶属度函数 (5) 和式 (3) 的计算得到 $\mu^k(\boldsymbol{X})$ (k=1, 2, …, K)；

④ 再经过式 (8) ~ (10) 的转化，得到映射后高维空间中的数据矩阵 $\boldsymbol{X}_g^{(1)} \in \boldsymbol{R}^{N\times(d+1)*K}$。

2) 多层递阶融合

① 利用 PCA 对 $\boldsymbol{X}_g^{(1)}$ 进行压缩，得到数据矩阵记为 $\boldsymbol{X}_{pca}^{(2)}$；

② For i=2:(S-1)；

③ 重复步骤 1)，对原数据进行模糊特征映射，得到数据矩阵 $\boldsymbol{X}_g^{(i)} \in \boldsymbol{R}^{N\times dK}$；

④ $\boldsymbol{X}_{lhffm}^{(i)}=\left[\boldsymbol{X}_{pca}^{(i)}\ \boldsymbol{X}_g^{(i)}\right]$；

⑤ 利用 PCA 对 $\boldsymbol{X}_{lhffm}^{(i)}$ 进行压缩，得到数据矩阵记为 $\boldsymbol{X}_{pca}^{(i+1)}$；

⑥ end；

2.2 基于多层递阶融合模糊特征映射的模糊 C 均值聚类算法 MLHFFFM-FCM

本节中，将多层递阶融合模糊特征映射与经典模糊聚类算法 FCM 相结合，提出基于多层递阶融合模糊特征映射的模糊 C 均值聚类算法。MLHFFFM-FCM 算法描述如下：

输入　给定一个数据集 $\boldsymbol{D}=\{\boldsymbol{X}, \boldsymbol{Y}\}$，设置初始模糊规则数 K，分层融合层数 S。

1) 通过基于多层递阶融合的模糊特征映射，将输入数据 $\boldsymbol{X}$ 转化为 $\boldsymbol{X}_{pca}^{(S)}$。

2) 对最终压缩融合获得的数据矩阵 $\boldsymbol{X}_{pca}^{(S)}$，采用 FCM 算法聚类。

输出　模糊划分矩阵 $\boldsymbol{U}$。

3 实验研究与分析

为了验证 MLHFFFM-FCM 算法在复杂非线性数据分析上的有效性，本节从 3 个方面进行对比分析：1) 各 FCM 演变算法之间聚类效果的对比实验；2) 单层映射结构与多层递阶融合映射结构的聚类效果对比实验；3) 关键参数敏感性的对比实验。

3.1 算法性能的评价指标

为了对各类算法的聚类性能进行对比，本文采用 NMI(normalized mutual information) 和 RI(rand

index) 作为实验评价指标。这两个指标的值越接近 1，说明算法聚类性能越好。其计算公式如下：

1) NMI

$$\mathrm{NMI}=\frac{\sum_{i=1}^{C}\sum_{j=1}^{C}N_{i,j}\log N\times N_{i,j}/N_i\times N_j}{\sqrt{\sum_{i=1}^{C}N_i\times\log N_i/N\times\sum_{j=1}^{C}N_j\times\log N_j/N}} \tag{14}$$

式中：$N_{i,j}$表示第i个聚类与第j类的契合程度，N_i表示第i个聚类所包含数据样本量，N_j表示类j所包含的数据样本量，而 N 表示整个数据样本大小。

2) RI

$$\mathrm{RI}=\frac{f_{00}+f_{11}}{N(N-1)/2} \tag{15}$$

式中：f_{00}表示数据点具有不同的类标签并且属于不同类的配对点数目，σ则表示数据点具有相同的类标签并且属于同一类的配对点数目，而 N 表示整个数据样本的总量大小。以上两种方法，其取值范围均为[0, 1]，且均随着数值的增大，显示出算法的性能更为优越。

3.2　实验设置

我们采用 UCI 真实数据集 (http://archive.ics.uci.edu/ml/) 来评估本文算法。为了测试实验应用数据集的广泛性以及避免选取数据集的偶然性，选择其中 7 个具有代表性的数据集 Ar_2、Diabetes、Zoo、Australian、Breast、Heart、Chronic_Kidney_Disease 进行测试，其中数据集的相关信息如表 1 所示。同时本文选取 5 种经典的聚类算法与 MLHFFFM-FCM 算法进行对比实验，分别为 FCM 算法、PCA-FCM 算法、ELM-FCM 算法、KFCM-K 算法以及 KFCM-F 算法。所有实验运行平台的配置如下：酷睿 i3 3.6 GHz CPU，3.42 G RAM，32 位 Windows 7 操作系统，MATLAB R2012b 编程环境。另外各算法相关说明及其参数设置如表 2 所示，其中各算法涉及的模糊指数 m 的寻优范围均为{1.2, 1.4, 1.6, 1.8, 2.0, 2.2, 2.4, 2.6, 2.8, 3.0, 3.2, 3.4, 3.6, 3.8, 4.0}。

表 1　实验数据集

Table 1　Experimental data sets

数据集	样本数 n	特征数 d	类别数 c
Ar_2	182	100	13
Diabetes	768	8	2
Zoo	101	16	7
Australian	690	14	2
Breast	277	9	2
Heart	270	13	2
Chronic_Kidney_Disease	400	24	2

表 2　各算法的说明以及相关参数设置

Table 2　The description of the algorithm and related parameters

算法	算法说明	相关参数	相关参数寻优范围设置
FCM[27]	模糊 C 均值聚类算法	模糊指数 m	
PCA-FCM	基于 PCA 特征提取的模糊 C 均值算法	模糊指数 m，特征提取数 d	d 的寻优范围为{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
ELM-FCM[28]	基于 ELM 隐空间映射的模糊 C 均值算法	模糊指数 m，隐节点数 n_h	n_h 的寻优范围为{100, 2, 00, 300, 400, 500, 600, 700, 800, 900, 1 000}
KFCM-K[29]	基于核空间的核模糊 C 均值聚类算法	模糊指数 m，核参数σ	σ由$\sigma^2=\frac{\sum_{i=1}^{n}\sum_{j=2}^{n}\lVert\boldsymbol{x}_i-\boldsymbol{x}_j\rVert^2}{n^2}$计算得出
KFCM-F[29]	基于特征空间的核模糊 C 均值聚类算法	模糊指数 m，核参数σ	σ由$\sigma^2=\frac{\sum_{i=1}^{n}\sum_{j=2}^{n}\lVert\boldsymbol{x}_i-\boldsymbol{x}_j\rVert^2}{n^2}$计算得出
MLHFFFM-FCM	基于多层递阶融合模糊特征映射的模糊 C 均值算法	模糊指数 m，特征提取数 d，高斯函数的宽度参数 h	d 的寻优范围从 1 到数据集本身维度的一半，h 的寻优范围为{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}}

3.3　聚类效果对比实验

为了验证 MLHFFFM-FCM 算法的有效性，本节对算法进行对比实验测试。在本实验中，将初始模糊规则数 r 设置为 30，多层递阶融合层数设置为 5 层，并根据表 2 的实验相关参数设置，分别对各算法重复运行 10 次。最终的实验中各算法的参数取值情况和实验结果如表 3 和表 4 所示。

表 3　各算法参数取值情况

Table 3　Parameter values of each algorithm

数据集	FCM	PCA-FCM	ELM-FCM	KFCM-K	KFCM-F	MLHFFFM-FCM
Ar_2	m=2.8	d=10 m=1.6	n_h=200 m=1.4	m=2.0 σ=0.8	m=3.4 σ=1.4	m=1.2 h=100 d=17
Diabetes	m=2.4	d=2 m=1.2	n_h=600 m=2.4	m=2.2 σ=0.4	m=1.6 σ=1.2	m=1.4 h=1 d=7
Zoo	m=1.4	d=5 m=1.4	n_h=1 000 m=1.8	m=1.2 σ=1.4	m=1.6 σ=0.8	m=1.8 h=100 d=9
Australian	m=2.0	d=2 m=2.6	n_h=100 m=2.0	m=1.2 σ=2.0	m=1.2 σ=1.4	m=1.2 h=100 d=9
Breast	m=3.6	d=1 m=1.2	n_h=100 m=1.2	m=1.2 σ=1.8	m=4.6 σ=0.4	m=1.4 h=10 d=1
Heart	m=3.2	d=3 m=3.2	n_h=100 m=2.2	m=4.6 σ=2.0	m=1.2 σ=0.6	m=2.6 h=10 d=2
Chronic_Kidney_Disease	m=3.6	d=7 m=4.0	n_h=800 m=1.2	m=1.2 σ=1.1	m=4.0 σ=1.1	m=2.6 h=10 d=11

表 4　各算法的运行结果

Table 4　Results of each algorithm

数据集	性能指标	FCM	PCA-FCM	ELM-FCM	KFCM-K	KFCM-F	MLHFFFM-FCM
Ar_2	RI_mean	0.772 5	0.950 1	0.860 5	0.866 4	0.838 6	0.979 0
	RI_std	0.050 5	0.002 5	0.005 0	0.006 7	0.019 8	0.001 1
	NMI_mean	0.565 6	0.789 3	0.459 7	0.333 7	0.627 1	0.935 3
	NMI_std	0.035 8	0.007 5	0.015 2	0.029 5	0.009 6	0.006 6
Diabetes	RI_mean	0.559 1	0.550 7	0.543 0	0.572 3	0.557 6	0.593 5
	RI_std	0.004 8	0	0.001 2	0.002 0	0	0.013 0
	NMI_mean	0.073 3	0.029 7	0.011 8	0.118 7	0.065 8	0.094 6
	NMI_std	0.008 3	0	0.004 4	0.003 0	0	0.020 1
Zoo	RI_mean	0.882 5	0.893 0	0.826 4	0.904 4	0.903 4	0.918 7
	RI_std	0.029 0	0.020 1	0.002 7	0.027 1	0.083 2	0.029 3
	NMI_mean	0.747 4	0.767 6	0.566 3	0.833 8	0.788 4	0.796 2
	NMI_std	0.035 0	0.028 8	0.005 8	0.021 1	0.111 0	0.024 5
Australian	RI_mean	0.728 5	0.507 1	0.505 0	0.743 6	0.733 6	0.753 9
	RI_std	0.080 2	0	0	0	0	0
	NMI_mean	0.388 0	0.034 4	0.009 9	0.415 9	0.399 2	0.431 0
	NMI_std	0.136 1	0	0	0	0	0
Breast	RI_mean	0.559 8	0.600 4	0.498 2	0.532 9	0.568 9	0.629 7
	RI_std	0.052 4	0	0.000 3	0.047 3	0.066 2	0.003 3
	NMI_mean	0.065 4	0.089 7	0.003 0	0.032 2	0.057 7	0.107 3
	NMI_std	0.051 0	0	0	0.042 2	0.056 7	0.007 8
Heart	RI_mean	0.522 9	0.522 9	0.504 8	0.667 4	0.683 3	0.737 3
	RI_std	0	0	0.003 0	0.002 4	0	0.011 7
	NMI_mean	0.032 8	0.032 8	0.038 8	0.260 9	0.280 6	0.387 0
	NMI_std	0	0	0.058 8	0.003 5	0	0.021 4
Chronic_Kidney_Disease	RI_mean	0.783 4	0.789 4	0.500 9	0.869 5	0.865 2	0.882 5
	RI_std	0.006 0	0.003 6	0	0	0	0
	NMI_mean	0.517 8	0.518 7	0.083 0	0.636 4	0.638 6	0.705 3
	NMI_std	0.008 6	0.005 3	0	0	0	0

从表 4 中可以明显地看出，在聚类精度上，文中涉及的对比算法只能在某个或某几个数据集上取得较优的结果，而 MLHFFFM-FCM 算法不仅在所有的测试数据集上取得满意的结果，并且还有着明显的提高。这说明了 MLHFFFM-FCM 算法的有效性，也进一步说明了该算法处理复杂非线性数据的强大能力。

3.4　单层映射结构与多层递阶融合映射结构的聚类效果对比实验与分析

为了体现本文算法引入的多层递阶融合方法的优越性，本节实验针对多层递阶融合映射结构对 FCM 算法性能的影响进行实验与分析。实验在模糊规则数设置相同的情况下，分别采用单层映射结构和多层递阶融合映射结构对原输入数据进行非线性映射，将映射后的数据采用 FCM 进行聚类。实验最终的参数取值情况和结果如表 5 和表 6 所示，其中因受篇幅所限，仅在表 6 中给出 RI 指标结果，NMI 与之有类似的结果，不再列出。

从表 5 和表 6 中可以明显地观察出，相比于单层映射结构，基于多层递阶融合映射结构的模糊聚类方法能够取得更好的学习效果。这是由于在单层映射之后的数据存在冗余信息，而在压缩之后又会导致信息缺失。但是多层递阶融合的映射结构是建立在单层映射结构的基础上，采用 PCA 技术对每一层模糊特征映射得到的高维特征表示进行压缩，再对应地结合每一层数据信息融合形成的。因此通过多层递阶融合的方法，可以有效地精简冗余信息，同时对每一层进行适当的信息弥补。这也充分体现了本文提出的多层递阶融合映射结构的优越。

表 5　两种算法结构的参数取值情况

Table 5　Parameter selection of two algorithms

算法映射结构	Ar_2	Diabetes	Zoo	Australian	Breast	Heart	Chronic_Kidney_Disease
单层	m=4.0	m=1.4	m=1.6	m=1.2	m=1.2	m=1.4	m=3.4
	h=100	h=1	h=100	h=100	h=0.1	h=10	h=10
多层递阶融合	m=1.2	m=1.4	m=1.8	m=1.2	m=1.4	m=2.6	m=2.6
	h=100	h=1	h=100	h=100	h=10	h=10	h=10
	d=17	d=7	d=9	d=9	d=1	d=2	d=11

表 6　两种算法结构的 RI_mean 性能指标

Table 6　Performance index of two algorithms

算法映射结构	Ar_2	Diabetes	Zoo	Australian	Breast	Heart	Chronic_Kidney_Disease
单层	0.859 0	0.568 0	0.894 5	0.753 9	0.597 1	0.704 1	0.876 2
多层递阶融合	0.979 0	0.593 5	0.918 7	0.753 9	0.629 7	0.737 3	0.882 5

3.5　参数敏感性实验

模糊规则数 r 作为 MLHFFFM-FCM 算法中的关键参数，本节针对该参数进行参数敏感性实验。这里为了让实验结果能够直观地进行观察与对比，我们同时对 KFCM-F 算法中的关键参数 σ 进行参数敏感性实验，进而研究模糊规则数这一关键参数对 MLHFFFM-FCM 算法性能的影响。实验中，MLHFFFM-FCM 模糊规则数 r 的实验取值范围为 {5, 10, 15, 20, 25, 30, 35, 40, 45, 50}，KFCM-F 算法中核参数 σ 的实验取值范围为 {0.1, 1.5, 10, 50, 100, 150, 200, 500, 1 000}，实验最终结果分别如图 2 和图 3 所示。

从图 2 中不难看出，KFCM-F 算法的性能随核参数 σ 变化出现很大的波动，这说明核参数 σ 对 KFCM-F 算法的性能有很大的影响。相反，由图 3 可以观察到，模糊规则数 r 对 MLHFFFM-FCM 算法性能的影响很小，算法性能始终保持稳定的状态，这说明 MLHFFFM-FCM 算法对模糊规则数 r 不敏感。结合上述实验也从另一个方面体现了采用本文提出的基于多层递阶融合映射方法的优越性，它不仅保证了算法的聚类效果，还克服了 KFCM-F 等算法对参数敏感的问题，这更有利于该算法在实际问题中的应用。

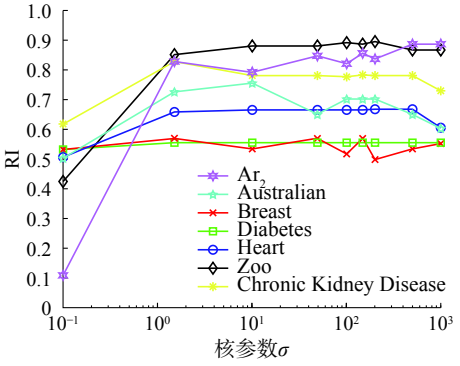

图 2　KFCM-F 算法性能随 σ 变化的影响

Fig. 2　Effect of σ on the performance of KFCM-F

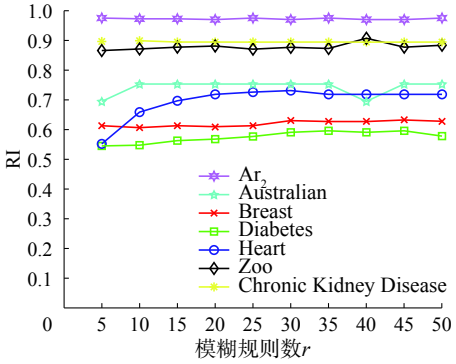

图 3　MLHFFFM-FCM 算法性能随模糊规则数 *r* 变化的影响

Fig. 3　Effect of fuzzy rules *r* on the performance of MLHFFFM-FCM

4　结束语

本文提出的 MLHFFFM-FCM 算法，是一种采用新型的显性映射方式来处理复杂非线性数据的无监督学习方法。相比于现有的核函数映射方法，MLHFFFM-FCM 算法在取得良好聚类效果的同时，还对算法中模糊规则数不敏感，这更有利于算法在实际应用中的选用。但是本文提出的 MLHFFFM-FCM 算法仍然具有一定的缺陷，例如对于高维数据，其时间开销较大。如何有效克服这些问题，将是今后进一步研究的重点。

参考文献：

[1] 王骏, 王士同, 邓赵红. 聚类分析研究中的若干问题[J]. 控制与决策, 2012, 27(3): 321–328.
WANG Jun, WANG Shitong, DENG Zhaohong. Survey on challenges in clustering analysis research[J]. Control and decision, 2012, 27(3): 321–328.

[2] 李宝刚. 基于读者日志分析的模糊聚类研究[J]. 价值工程, 2011, 30(33): 146–147.
Li Baogang. The fuzzy clustering on analyzing reader's log[J]. Value engineering, 2011, 30(33): 146–147.

[3] PENG Hong, WANG Jun, PÉREZ-JIMÉNEZ M J, et al. An unsupervised learning algorithm for membrane computing[J]. Information sciences, 2015, 304: 80–91.

[4] QIN Chen, SONG Shiji, HUANG Gao, et al. Unsupervised neighborhood component analysis for clustering[J]. Neurocomputing, 2015, 168: 609–617.

[5] XU Yan, QIU Peng, ROYSAM B. Unsupervised discovery of subspace trends[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(10): 2131–2145.

[6] 杨玉梅. 基于信息熵改进的 K-means 动态聚类算法[J]. 重庆邮电大学学报: 自然科学版, 2016, 28(2): 254–259.
YANG Yumei. Improved K-means dynamic clustering algorithm based on information entropy[J]. Journal of Chongqing university of posts and telecommunications: natural science edition, 2016, 28(2): 254–259.

[7] 阎辉, 张学工, 李衍达. 基于核函数的最大间隔聚类算法[J]. 清华大学学报: 自然科学版, 2002, 42(1): 132–134.
YAN Hui, ZHANG Xuegong, LI Yanda. Kernel-based maximal-margin clustring algorithm[J]. Journal of Tsinghua university: science and technology, 2002, 42(1): 132–134.

[8] MA Bo, QU Huiyang, WONG H S. Kernel clustering-based discriminant analysis[J]. Pattern recognition, 2007, 40(1): 324–327.

[9] LIAO Li, ZHOU Jianzhong, ZOU Qiang. Weighted fuzzy kernel-clustering algorithm with adaptive differential evolution and its application on flood classification[J]. Natural hazards, 2013, 69(1): 279–293.

[10] 李侃, 刘玉树. 模糊核聚类的自适应算法[J]. 控制与决策, 2004, 19(5): 595–597.
LI Kan, LIU Yushu. Fuzzy kernel clustering self-adaptive algorithm[J]. Control and decision, 2004, 19(5): 595–597.

[11] WANG Jun, DENG Zhaohong, JIANG Yizhang, et al. Multiple-kernel based soft subspace fuzzy clustering [C]//Proceedings of 2014 IEEE International Conference on Fuzzy Systems. Beijing, China, 2014: 186–193.

[12] WANG Jun, DENG Zhaohong, CHOI K S, et al. Distance metric learning for soft subspace clustering in composite Kernel space[J]. Pattern recognition, 2015, 52: 113–134.

[13] GIROLAMI M. Mercer kernel-based clustering in feature space[J]. IEEE transactions on neural networks, 2002, 13(3): 780–784.

[14] MÉNDEZ G M, DE LOS ANGELES HERNÁNDEZ M. Hybrid learning mechanism for interval A2-C1 type-2 non-singleton type-2 Takagi-Sugeno-Kang fuzzy logic systems[J]. Information sciences, 2013, 220: 149–169.

[15] TSAKONAS A, GABRYS B. Evolving Takagi-Sugeno-Kang fuzzy systems using multi[J]. Journal of clinical endocrinology and metabolism, 2011, 96(12): 3603–3608.

[16] CHUANG C C, SU Shunfeng, CHEN S S. Robust TSK fuzzy modeling for function approximation with outliers [J]. IEEE transactions on fuzzy systems, 2001, 9(6): 810–821.

[17] SUGENO M, KANG G T. Structure identification of fuzzy model[J]. Fuzzy sets and systems, 1988, 28(1): 15–33.

[18] PRICE A L, PATTERSON N J, PLENGE R M, et al. Principal components analysis corrects for stratification in genome-wide association studies[J]. Nature genetics, 2006, 38(8): 904–909.

[19] JOLLIFFE I T. Principal component analysis[M]. Berlin: Springer, 2012: 41–64.

[20] 冯斌, 须文波. 基于 TSK 模糊系统的生化变量预估模型

[J]. 计算机与应用化学, 2006, 23(4): 343–346.
FENG Bin, XU Wenbo. Biochemical variable estimation model based on TSK fuzzy system[J]. Computers and applied chemistry, 2006, 23(4): 343–346.

[21] WU Dongrui. Approaches for reducing the computational cost of interval type-2 fuzzy logic systems: overview and comparisons[J]. IEEE transactions on fuzzy systems, 2013, 21(1): 80–99.

[22] DENG Zhaohong, CHOI K S, CHUNG F L, et al. Scalable TSK fuzzy modeling for very large datasets using minimal-enclosing-ball approximation[J]. IEEE transactions on fuzzy systems, 2011, 19(2): 210–226.

[23] 蒋亦樟, 邓赵红, 王士同. ML 型迁移学习模糊系统[J]. 自动化学报, 2012, 38(9): 1393–1409.
JIANG Yizhang, DENG Zhaohong, WANG Shitong. Mamdani-larsen type transfer learning fuzzy system[J]. Acta automatica sinica, 2012, 38(9): 1393–1409.

[24] LESKI J M. TSK-fuzzy modeling based on ε-insensitive learning[J]. IEEE transactions on fuzzy systems, 2005, 13(2): 181–193.

[25] ZHOU Hongming, HUANG Guangbin, LIN Zhiping, et al. Stacked extreme learning machines[J]. IEEE transactions on cybernetics, 2015, 45(9): 2013–2025.

[26] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436–444.

[27] BEZDEK J C, EHRLICH R, FULL W. FCM: the fuzzy c-means clustering algorithm[J]. Computers and geosciences, 1984, 10(2/3): 191–203.

[28] HE Qing, JIN Xin, DU Changying, et al. Clustering in extreme learning machine feature space[J]. Neurocomputing, 2014, 128: 88–95.

[29] GRAVES D, PEDRYCZ W. Kernel-based fuzzy clustering and fuzzy clustering: a comparative experimental study[J]. Fuzzy sets and systems, 2010, 161(4): 522–543.

作者简介：

鲍国强，男，1992 年生，硕士研究生，主要研究方向为智能计算与模式识别。

应文豪，男，1979 年生，副教授，博士，主要研究方向为模式识别与智能计算。

蒋亦樟，男，1988 年生，讲师，博士，主要研究方向为模式识别与智能计算。